DOCUMENT RESUME

ED 477 932 TM 035 028

AUTHOR	Schwarz, Richard D.; Rich, Changhua; Podrabsky, Tracy
TITLE	A DIF Analysis of Item-Level Mode Effects for Computerized and Paper-and-Pencil Tests.
PUB DATE	2003-04-00
NOTE	32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Adult Basic Education; Adults; *Computer Assisted Testing; Elementary School Students; Elementary Secondary Education; *Item Bias; Nonparametric Statistics; Secondary School Students; *Test Format; Testing Problems; *Validity
IDENTIFIERS	Item Bias Detection; *Standardized Mean Difference

ABSTRACT

This paper studied the usefulness of differential item functioning (DIF) methodology for examining potential mode effects. Although the goal was not to validate the comparability of the assessments per se, it is of interest to speculate why some formats could give rise to differential performance. Data were obtained from two instruments on which validation studies were conducted to determine if scores were comparable across modes of administration. The first instrument was a norm-referenced aptitude test called InView, administered to students in grades 4 to 9 in computer and paper modes for this study. Additional data were obtained from a Test of Adult Basic Education (TABE), also norm-referenced. Test takers were asked about their levels of computer experience and preferences. Two DIF statistics, the Linn-Harnisch procedure (1981) and the nonparametric Standardized Mean Difference were used to assess DIF. Results show that some levels of InView had substantial numbers of items that were flagged for the online comparisons with the standardization study, while others had relatively few. The TABE had two flagged items, and differences between computer-based and standardization groups were throughout the ability range with more apparent differences in the lower portion of the distribution. In general, students were neutral to mode or preferred computer-administered tests. Findings show that DIF methodology presents a well-studied method for examining group differences at the item level that can be used to examine mode of administration differences at the item level. Appendixes contain the student surveys and derivations of the DIF methods. (Contains 10 figures, 4 tables, and 10 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
C. Rich
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

ED 477 932

A DIF Analysis of Item-Level Mode Effects for Computerized and Paper- and -Pencil Tests

by

Richard D. Schwarz
Changhua Rich
Tracy Podrabsky
CTB/McGraw-Hill

TM035028

Presented at NCME
Chicago, April 2003

We would like to acknowledge Ping Wan, Cristina Hangakoon, Bob Kelly, and Raylene Potter for their programming assistance.

BEST COPY AVAILABLE

Increasingly, computer-based assessments are being administered in grades 2 to 12. Many of these computerized tests have a legacy in paper- and -pencil administrations such that it is necessary to establish the comparability of scores across these modes (Mead & Drasgow, 1993). As a goal, the scores emanating from each test should be comparable across modes of administration. Comparability can be defined as the differences between modes of administration that are equivalent to those observed between alternate (parallel) forms of the test. It would be very convenient if scores could be compared across administrations without resorting to some type of score transformation (i e., equating). However, comparable scores are sometimes difficult to achieve between paper- and - pencil and linear test administrations. For instance, the usability of computer administered tests in Reading could differ from paper- and -pencil versions when younger students are required to scroll through a reading passage in order to both read it and respond to questions related to that passage. For younger students, perceptual differences in the display of information could potentially lead to disparities in performance (Choi & Tinker, 2000). Differences across modes of administration might also be evident with adults that have low levels of functional literacy. Most research to date has focused on college bound students and college graduates (Schaeffer, Reese, Steffen, McKinley & Mills, 1993). Other sources of mode effects might emerge for school age students or adults in alternative educational settings.

Cross-mode comparability has traditionally been assessed by a variety of methods such as correlations, differences in scale scores, or effect sizes. Comparison of overall test statistics will not reveal which items are sensitive to mode of administration. This paper proposes the use of differential item functioning methodology (DIF) to examine differences in performance across modes of test administration. There are a number of advantages to using this DIF approach. First, DIF is routinely used in testing programs to examine potential performance differences among various demographic groups such as gender or ethnicity. This methodology can easily be redeployed to examine group differences across mode of test administration. Depending on the nature of the comparability study, one mode such as paper- and -pencil could be designated as the reference group and the other mode as the focal group. Using DIF procedures permits the examination of differences between modes of administration at the level of the item. This would be particularly important in determining the precise sources of noncomparablitiy between modes of administration were they to occur. DIF methodology also provides statistical tests that can be used to flag items for further examination. Surveys suggest that students perceived scrolling as the biggest problem with computerized assessments of reading ability (Swanson, 2002). DIF analysis can confirm what types of formats lead to significant differences in performance across modes of administration. In order to increase the generalizability of these findings and demonstrate the utility of DIF methods, information from two different types of assessments were utilized in order to investigate the effect of different item formats with varying grade levels. Since these assessments were both norm-referenced, student performance on computer-based tests will be compared with standardization samples that used pencil and paper. Within this context, the primary objective of these comparability studies is to establish that the on-line scores

are valid and defensible for the intended purposes. The intent of this paper is to demonstrate the utility of DIF methodology for examining potential mode effects. The goal was not to validate the comparability of these assessments, per say. However, it will be of some interest to conjecture why some formats could give rise to differential performance.

Method

Instruments

Data was obtained from two instruments in which validation studies were conducted in order to determine if scores are comparable across modes of administration. National standardization studies in which norms were derived for paper administrations were carried out for these instruments. The first instrument a norm-referenced aptitude test called InView, which contains Sequences, Analogies, Quantitative Reasoning, and Verbal Reasoning (Words & Context) subtests. Each subtest consisted of 20 selected-response items that contain different types of item formats. InView is administered in a cross grade configuration (i.e., multiple grades per test level). For purposes of this study, grades 4 to 9 were examined. The computer-based administration differed from the paper and pencil administration in that a single item was presented on-screen at a time rather than 4 to 6 being present on a single printed page in the test book. Items were linearly administered in the same order as pencil- and –paper. Students had the opportunity to go back to an item and change their response as necessary. The Analogies and Quantitative Reasoning was the focus of these investigations primarily due to the greater variation and complexity in item formats present compared with other InView subtests. The Analogies subtest is a nonverbal measure of student's ability to discern various types of relationships among picture pairs and infer parallel relationships between incomplete ones. The Quantitative Reasoning subtest is intended to measure student's ability to think with numbers and solve problems through complex quantitative reasoning processes. A survey was administered to the examinees that took the computer-based tests (Appendix A). The survey was used to collect information concerning examinees' familiarity with computers and their perceptions of computer administered tests.

Additional data was obtained from a Test of Adult Basic Education (TABE), which is a norm-referenced instrument that measures basic concepts and knowledge in Reading, Mathematics, and Language using selected-response items. TABE was normed in paper- and –pencil mode, and given in a linear computer-based mode for the comparability study. TABE has five levels L, E, M, D, A (Limited literacy, Easy, Medium, Difficult, and Advanced). Note that TABE levels correspond to content grade range of 0.0 to 12.9. Level M corresponds to grade range of 4-5, indicating relatively low levels of adult literacy. Most items in TABE Reading are associated with passages. Many passages require scrolling of the text in order to read them. A survey was administered to the examinees to collect information concerning examinees' familiarity with computers and their perceptions of computer administered tests (Appendix A).

Samples

The sample composition for InView and TABE respectively are shown in Tables 1 and 2 for the standardization and the on-line sample. Both InView and TABE were calibrated and normed using nationally representative samples in traditional paper and pencil administrations. The InView comparability study occurred subsequent to the national standardization study. In this study, a heterogeneous sample based on region, city size, and correlates of achievement (free & reduced lunch) were selected for the on-line administrations. A large number of students took the computer-based administration in the event that the on-line version needed to be calibrated and equated to achieve comparability with the standardization sample (pencil and paper). For the TABE comparability study, a subset of students took the computer-based test concurrently with the national standardization sample. It should be noted that TABE is administered to adults in alternative educational settings. As a result, this population is largely male with a high proportion of African-American and Hispanic students. Table 2 shows that for the computer-based administration larger numbers of females were present in the sample compared with standardization.

Descriptive Statistics

Tables 3 and 4 respectively show descriptive statistics for InView and TABE. Table 3 shows InView raw- and scale score statistics for the comparability study and the national standardization. Since the IRT estimates for the comparability study were on the logistic (0,1) metric, the linear transformation constants from the national standardization were applied to them. The on-line mean scores were substantially lower than either of the paper- and pencil administrations for the InView tests. For example, the mean raw score of 12.3 out of twenty possible points for Level 2 (grades 4 & 5) Analogies was lower than the grade 4 mean raw score from standardization of 13.7. These performance differences may be due to higher omit rates for students taking the test on-line. For this study, students may have skipped items, failed to respond, or the system may have failed adequately to capture them. Table 4 shows the descriptive statistics for the TABE standardization and the computer administration. The mean raw scores and standard deviation are comparable between the two modes for Reading test of Level M and Level D.

Procedure

Scaling. Selected-response items for both instruments were scaled using the three-parameter logistic model (Lord & Novick, 1968; Lord, 1980) in which the probability that a student with ability θ responded correctly to item i is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}$$

where a_i denotes the item discrimination, b_i the item difficulty, and c_i the lower asymptote corresponding to the probability of a correct response by a very low-scoring student. The three-parameter model was estimated using marginal maximum likelihood

procedures (Bock & Aitkin, 1981) via the IRT scaling program PARMATE (Burket, 1991). Scale scores were derived using maximum likelihood that have an arbitrary linear transformation applied.

DIF. The modes of administration were used as a grouping variable for the DIF analysis where the standardization samples were used as the reference group. DIF analyses were conducted prior to any equating that could be used to achieve comparability between modes. Two DIF statistics, the Linn-Harnisch procedure (1981) and the more familiar non-parametric Standardized Mean Difference (Zwick, Donoghue, and Grima, 1993) were used for assessing DIF. Other choices of DIF statistics could have also been made. The Linn-Harnisch procedure was used primarily since it gives results based on the operational IRT score metric. In the Linn-Harnisch analyses, the observed proportion correct of the computer administration can be compared with the expected proportion correct estimated using the paper- and -pencil sample. Standardized Mean Difference (SMD) was used primarily as a cross-validation technique. Likewise in the SMD analyses, computer mode can be compared with the reference subgroup that was administered the items in paper- and -pencil mode. The students who were administered the paper- and -pencil administrations were designated as the reference groups in the DIF analysis since the norms for these tests were derived from the paper- and -pencil administration for both TABE and InView.

For the Linn-Harnisch procedure, an item is flagged for DIF (against or favoring a subgroup) when the observed minus the expected mean proportion correct is greater or equal than the absolute value of 0.10 and the corresponding Z value is greater than or equal to the absolute values of 2.58. The SMD index expresses results in an item-level score metric. The mean SMD of - .11 indicates that on the average, the difference in mean item score (focal - reference) is more than 1/10 of a score point (Zwick, Donoghue, & Grima, 1993). SMD with an absolute value of .10 and larger was flagged for DIF, that is, the items demonstrating 1/10 of a score point difference between focal and reference group comparison. The purpose for this criterion is to identify item formats with substantial sensitivity to mode effects. From an exploratory standpoint, there is particular interest in items that were flagged by both statistics. The derivation for both these DIF procedures is given in Appendix B. Differences in item characteristic curves for some flagged items were evaluated graphically across the ability distribution.

Results

Tables 5 and 6 show the results of the DIF analyses for InView and TABE. A negative Standardized Mean Difference or Linn-Harnisch indicates that the items were flagged against the standardization (paper-pencil) group. Table 5 shows that some levels of InView had substantial numbers of items that were flagged for the on-line comparisons with the standardization study while others had relatively few. For Level 3 of InView, two items were flagged against the standardization reference group and two for items the focal on-line group resulting in DIF cancellation for this subtest. Analogies items B (Level 3) and F (Level 4) had large differences for both DIF statistics. Table 6 shows the

items in TABE level M and level D Reading that were flagged for DIF. Notice that all of the items flagged had small to moderate DIF. All items flagged in the Reading test required scrolling the passages in the computer-based test. Although each reading passage was associated with four or more selected-response items, only one of them was flagged for DIF. The specific information needed to answer these flagged items may have required more scrolling back and forth of the passage in order to respond. Table 6 indicates if scrolling was a feature present in the passage.

Using the results from Table 5 and 6 an examination of the operating characteristics for several items formats are given in Figures 1, 2, 3, 4 and 5 using the national standardization as the reference group. The smoothed model derived probability of getting the item correct for a given ability level using the standardization data is plotted for these items. The bubble plot below shows the density of students in a given ability region. Figure 1, Analogies Level 3 item B, shows that this item was easier for on-line students throughout most the ability distribution. Students within the standardization sample had to have higher scale scores in order to have the same probability of getting the item correct compared with the on-line group. This item required the students to discern the analogy between an untrimmed bush and a trimmed one with a shaggy and a shaved poodle. Answering this item required a fine visual discrimination to be made on the part of the student. The on-line stimulus was slightly larger than the one in the printed book and may have accounted for this difference. Figure 2 shows the operating characteristics for Item G of Quantitative Reasoning Level 3, which was harder for the on-line group. This item required a comparison of the irregular shaded areas of a grid and the selection of the one with the greatest area shaded. Comparison of area may be more difficult on a computer screen than on paper. Figure 3 shows the operating characteristics for Item F of Analogies Level 4, which was harder for the on-line group. This is the opposite of the effect seen for the Level 3, item B in Figure 1 in which students also had to make a fine visual discrimination of pictorial content.

Figures 4 and 5 show the two flagged items for TABE Level M Reading for Items J and L. Both of these items required scrolling. The differences between the computer-based and standardization groups were throughout the ability range with more apparent differences in the lower portion of the distribution.

Figures 6 to 10 show the results of surveys given to students at each respective test level for InView and TABE regarding their computer experiences and preferences. Question numbers two and six on the InView Survey asked students about their preference for tests administered on computer versus paper- and -pencil tests. Most students were either neutral to mode or preferred tests administered on the computer. For the TABE survey, question three asked students about their experience taking TABE on the computer. Since the TABE survey was not a likert scale, the survey will need to be referenced in order to understand the categories being graphed in Figures 9 and 10. The preponderance of students indicated that they do their best work on a computer and that they are comfortable with this format. Question five on the TABE survey asked students about scrolling passages. Most students indicated that it was easier or about the same as reading a paper- and –pencil test.

Discussion

The Standards for Educational and Psychological Testing (4.10) suggest that a clear rationale and supporting evidence be supplied for any claim that scores earned on different forms or modes of a test are interchangeable. DIF methodology presents a powerful and well-studied method for examining group differences at the item level that is used to augment validity arguments. This methodology can easily be applied to the issue of examining mode of administration differences at the item level. DIF methodology allows potential mode differences to be identified using a statistical criterion.

For gender and ethnicity groups, the current practice is to iteratively purify the test of DIF by eliminating items with differential operating characteristics from the test selection. Items displaying DIF across mode would be far less likely to be eliminated from a selection due to many other content or psychometric requirements. For instance, the content validity of a test would be threatened if Reading passages that required scrolling were eliminated. If strong mode effects were identified, then equating methods could be used to achieve comparability. In testing programs in which both computerized and paper- and -pencil administrations are both used interchangeably, the test blueprint could specify item formats that lend themselves more easily to comparability across modes. Item formats could be devised that tend to minimize mode effects while not threatening the content validity of the test. Item-level information derived from DIF analyses could be used by test developers to evaluate and design items that minimize differences in performance across modes of administration. Differences that occur throughout the ability distribution might reflect additional task processing demands for computerized assessments or the need for additional tutorials. It is probable that no suggestions for changes to these item formats would have been made based on the results of this study. However, to ensure comparability the recommendation would be to ensure that item formats and features are as similar as possible with the paper- and -pencil mode since these are the source of the normative inferences. For instance, good practice would dictate that items have the same size and shading across test formats. DIF methodology affords some additional methods for examining these types of questions at the item level that can further inform test construction in these types of programs.

From an exploratory standpoint, it should be noted that the item types contained in these two instruments had relatively simple selected-response formats. Simple selected-response formats can be defined as items that are discrete (not locally dependent) and lack more complicated graphical elements. Items with simple formats might be less likely to give rise to mode of administration effects. Selected-response items that have more complex formats contain more complicated graphical elements and entail objects that need to be manipulated by the student in order to respond. For instance, an item might require the length of an object to be measured. Measuring the length of an object on a screen using a toolbox maybe a very different construct compared with the use of a

ruler and a test booklet for school age students. More mode DIF might be expected for items with more complicated formats.

The differences in performance across mode for InView deserves some comment. One expectation is that many items should have been flagged for DIF based on an inspection of the mean differences between modes of administration shown in Table 1. For example, Level 2 of InView had substantial differences in scores distributions but had few items flagged for DIF. This is similar to the investigation of ethnic DIF when the distribution of ability across groups is substantially different. These differences in ability distributions do not necessarily give rise to DIF since these statistics are conditional on ability.

Finally, student surveys addressing the issues of accessibility, familiarity, comfort level/anxiety, scrolling effect and tutorial effectiveness provide important sources of validity evidence. The two surveys seemed to indicate that students were comfortable with computer-based test administrations. One area not well addressed in this paper is to examine differential performance based on particular responses to survey questions as a grouping variable. That is, how do student that indicate "Taking the test on a computer was easier" perform both in absolute terms and conditional on ability?

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443–459.

Burket, G. R. (1991; 1995). *PARMATE* [Computer program]. Unpublished.

Choi, S.W. & Tinkler, T. (April, 2002). Evaluating comparability of paper-and pencil and computerized-based assessments in a K-12 setting. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (71, 179–181). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mead, A.D. & Drasgow, F. (1993). Equivalence of computerized and paper-and Pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Schaeffer, G.A., Reese, C.M., Steffen, M., McKinley, R.L., & Mills C.N. (1993). Field test of a computerized-based GRE General Test ETS Research Report 93-07.

Swanson, L. (2002). Issues in transitioning state assessments to computer delivery. Paper presented at the CCSSO 2002 Conference on Large-Scale Assessment.

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

Table 1.
Sample characteristics for InView national standardization and on-line samples

National Standardization Sample (Paper- and -Pencil)

Test				Ethnicity			Gender	
Level	Grade	Subtest	N	African-American	Hispanic	Other[+]	Female	Male
2	4 & 5	AN	10301/ 9653	1492/1162	805/ 906	4645/5080	3433/3606	3466/3491
3	6 & 7	AN	10202/10232	1455/1561	1077/ 979	5922/4761	4251/3510	4084/3660
4	8 & 9	AN	9611/11034	1358/1746	1221/1062	3942/4815	3261/3925	3068/3502
2	4 & 5	QR	10250/ 9654	1492/1162	805/ 906	4645/5080	3433/3606	3466/3491
3	6 & 7	QR	10149/10179	1455/1561	1077/ 979	5922/4761	4251/3510	4084/3660
4	8 & 9	QR	9853/10832	1358/1746	1221/1062	3942/4815	3261/3925	3068/3502

On-Line Sample

Test				Ethnicity			Gender		
Level	Grade	Subtest	N	African-American	Hispanic	Other	Female	Male	Omitted
2	4 & 5	AN	2295	587	250	1458	1106	1050	139
3	6 & 7	AN	1839	200	426	1213	846	886	107
4	8 & 9	AN	1455	183	202	1070	746	655	54
2	4 & 5	QR	2103	514	244	1345	1030	952	121
3	6 & 7	QR	1623	157	389	1077	749	775	99
4	8 & 9	QR	1260	142	193	925	651	557	52

Note: *AN refers to Analogies and QR to Quantitative Reasoning. [+]Other includes Caucasians.

Table 2.
Sample characteristics for TABE national standardization and computer-based samples for Reading

National Standardization Sample (Paper- and -Pencil)

Test		Ethnicity			Gender		
Level	N	African-American	Hispanic	Other*	Female	Male	Omitted
M	1522	689	352	494	375	1119	41
D	1690	730	233	731	389	1278	27

Computer-based Sample

Test		Ethnicity			Gender		
Level	N	African-American	Hispanic	Other	Female	Male	Omitted
M	207	72	38	97	80	127	0
D	219	119	29	71	121	98	0

Note: *Other includes Caucasians.

Table 3.
InView descriptive statistics for two modes of administration

National Standardization (Paper)

Test Level	Grades	Subtest	N by grade	Raw score by grade		Scale Score by grade	
				Mean	SD	Mean	SD
2	4 & 5	AN	10301/ 9653	13.7/14.8	4.0/3.7	441/460	65/62
3	6 & 7	AN	10202/10232	13.2/13.8	4.2/4.1	479/488	64/64
4	8 & 9	AN	9611/11034	12.9/13.3	4.4/4.4	502/507	59/59
2	4 & 5	QR	10250/ 9654	12.1/13.2	4.2/4.2	453/467	52/53
3	6 & 7	QR	10149/10179	12.1/12.7	4.0/4.2	483/493	59/62
4	8 & 9	QR	9853/10832	11.6/12.0	4.6/4.7	506/512	62/65

On-Line

Test Level	Grades	Subtest	N	Raw Score		Scale Score	
				Mean	SD	Mean	SD
2	4 & 5	AN	2295	12.3	4.2	417	72
3	6 & 7	AN	1839	11.4	4.1	447	68
4	8 & 9	AN	1455	11.0	4.5	473	63
2	4 & 5	QR	2103	9.8	4.4	417	61
3	6 & 7	QR	1623	10.2	4.0	452	67
4	8 & 9	QR	1260	10.2	4.3	485	67

Note: *AN refers to Analogies and QR to Quantitative Reasoning.

Table 4.
TABE descriptive statistics for two modes of administration for Reading

National Standardization (Paper)					
Test		Raw Score		Scale Score	
Level	N	Mean	SD	Mean	SD
M	1522	35.4	10.5	510.1	71.7
D	1690	32.0	9.8	518.7	74.9

Computer-based					
Test		Raw Score		Scale Score	
Level	N	Mean	SD	Mean	SD
M	207	34.8	9.1	508.7	62.6
D	219	33.1	10.0	527.6	76.1

Table 5.
Items flagged for DIF for InView Analogies and Quantitative Reasoning

Test Level	Item	Standardized Mean Difference	Linn-Harnisch	
		Analogies	Difference	Z
2	A	-.11	-.12	-13.1
3	B	+.20	+.20	+19.8
	C	+.10	+.11	+10.4
	D	-.10	-.10	-9.5
	E	-.12	-.12	-13.8
4	F	-.21	-.21	-29.2
		Quantitative Reasoning		
2	-	No items flagged	No items flagged	No items flagged
3	G	-.14	-.16	-13.1
4	H	-.12	-.13	-10.2

Table 6.
Items flagged for DIF for TABE Reading

Test Level	Item	Scrolling Present	Standardized Mean Difference	Linn-Harnisch	
				Difference	Z
M	I	No	+0.11	+0.12	+3.8
M	J	Yes	-0.13	-0.15	-5.1
M	K	Yes	-0.09	-0.11	-3.6
M	L	Yes	-0.12	-0.13	-3.9
D	M	Yes	-0.10	-0.11	-5.5
D	N	Yes	-0.10	-0.12	-4.4
D	O	Yes	-0.10	-0.13	-4.3
D	P	No	+0.11	+0.10	+2.8

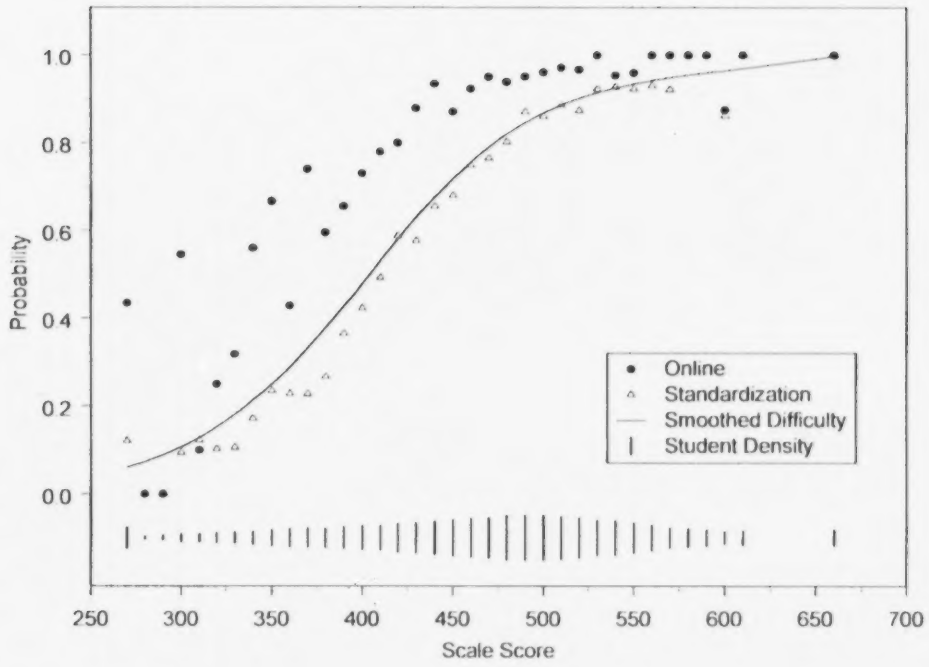

Figure 1.
Item characteristic curve for Analogies Level 3, item B

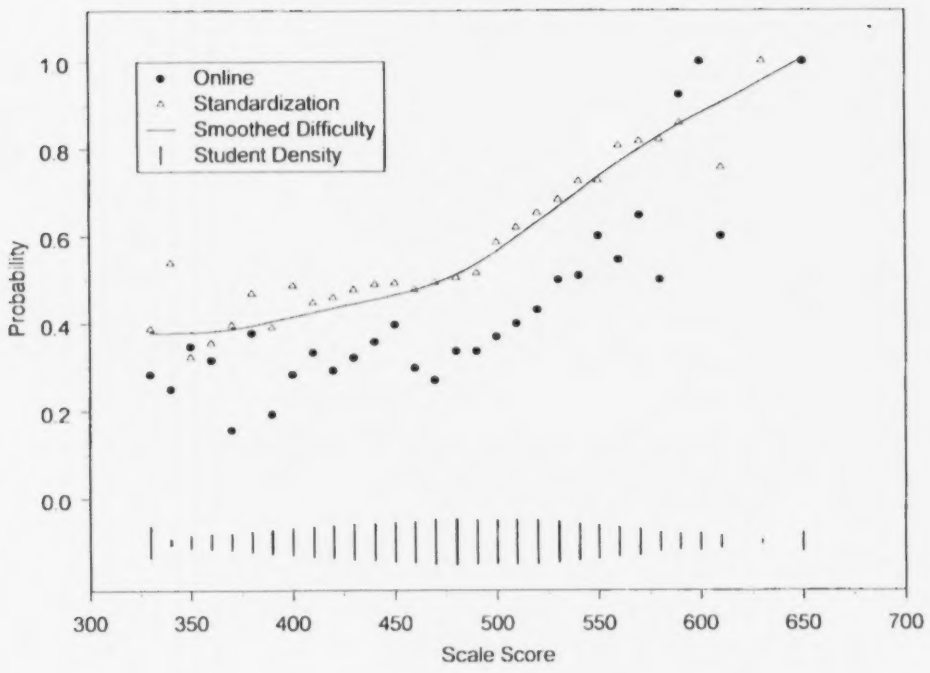

Figure 2.
Item characteristic curve for Quantitative Reasoning, Level 3 item G

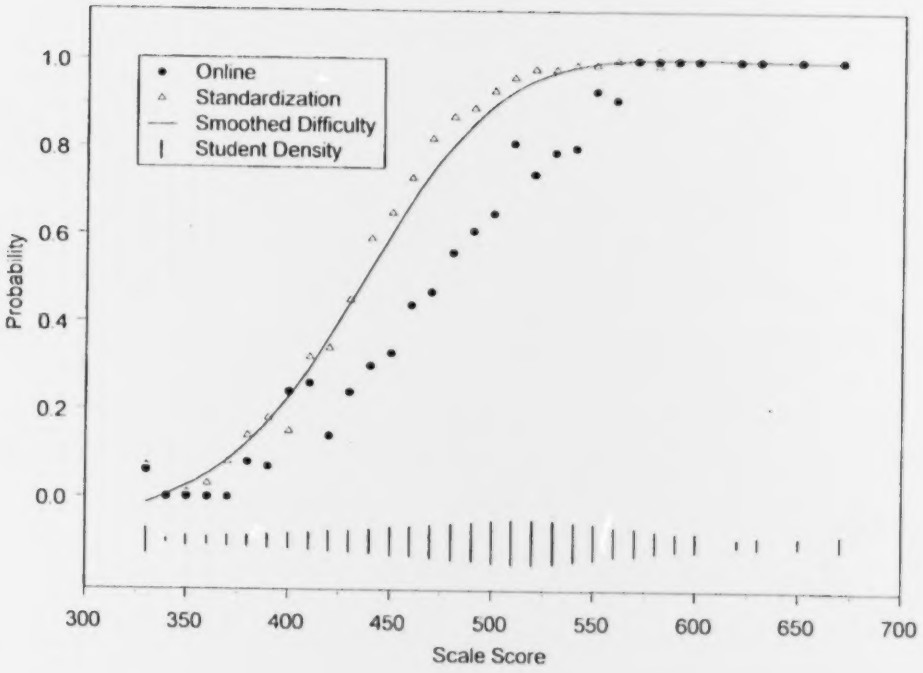

Figure 3.
Item characteristic curve for Analogies Level 4, item F

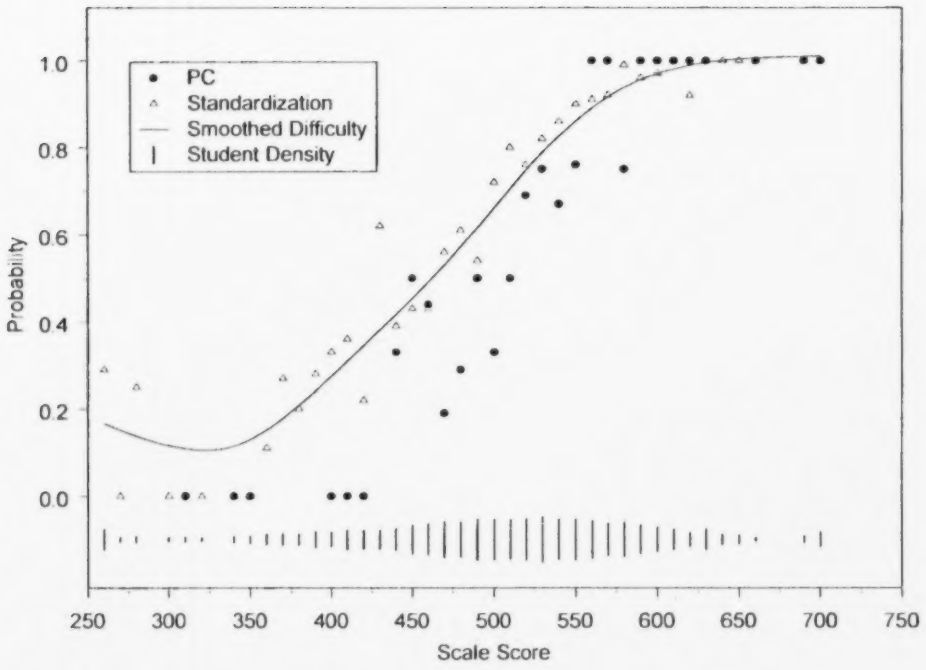

Figure 4.
Item characteristic curve for TABE Level M, item J

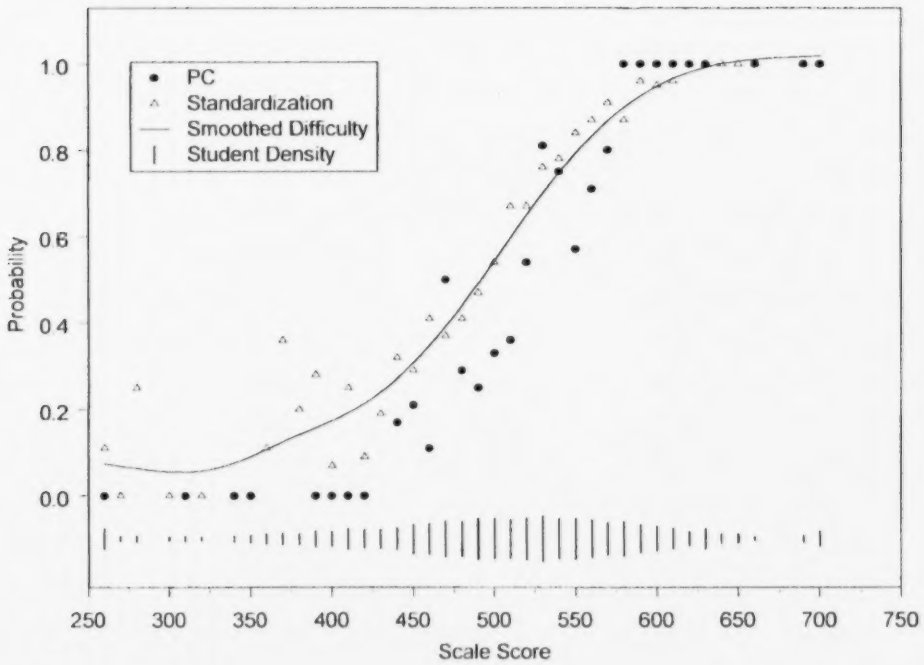

Figure 5.
Item characteristic curve for TABE Level M, item L

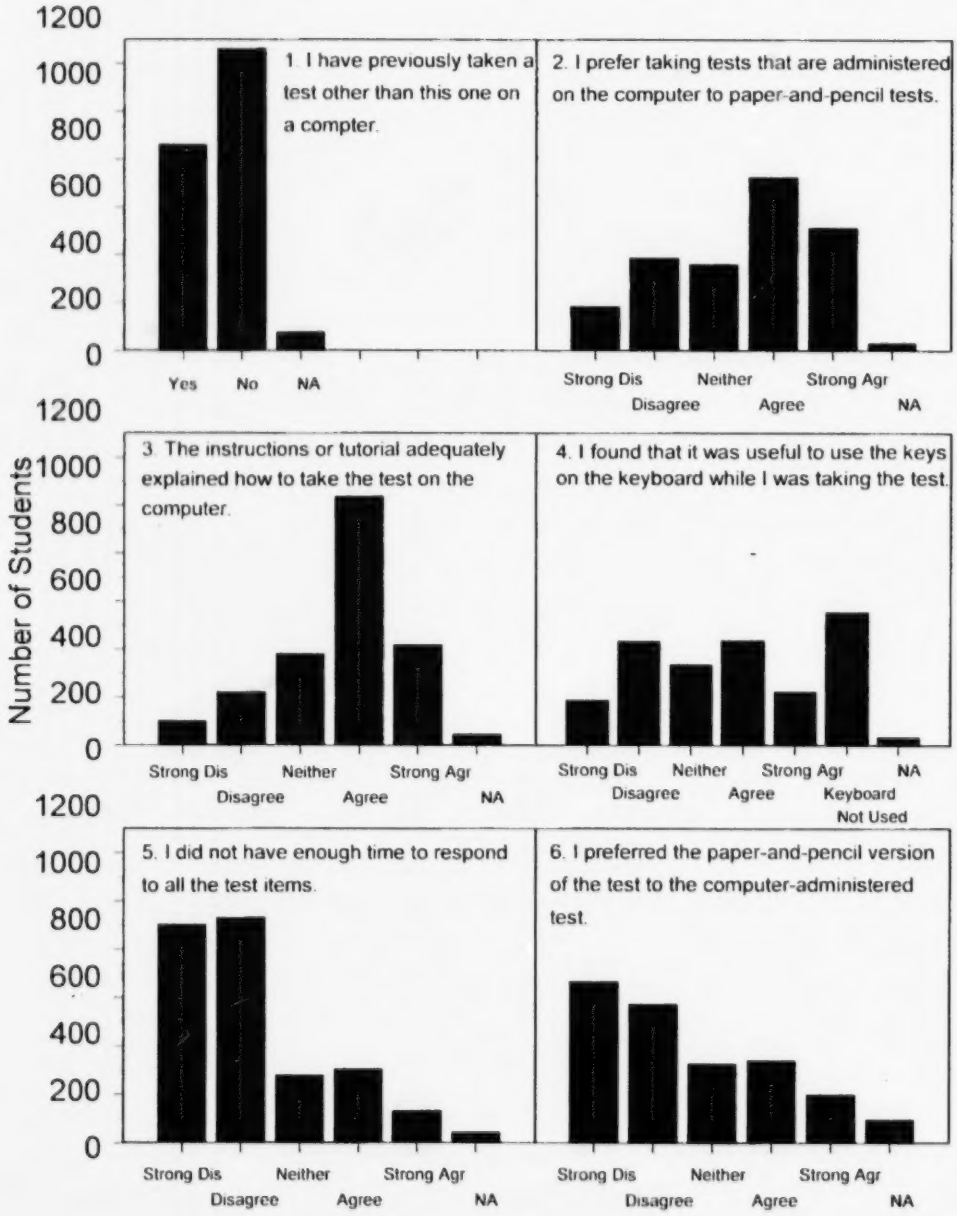

Figure 6.
Responses to survey questions for InView Level 2

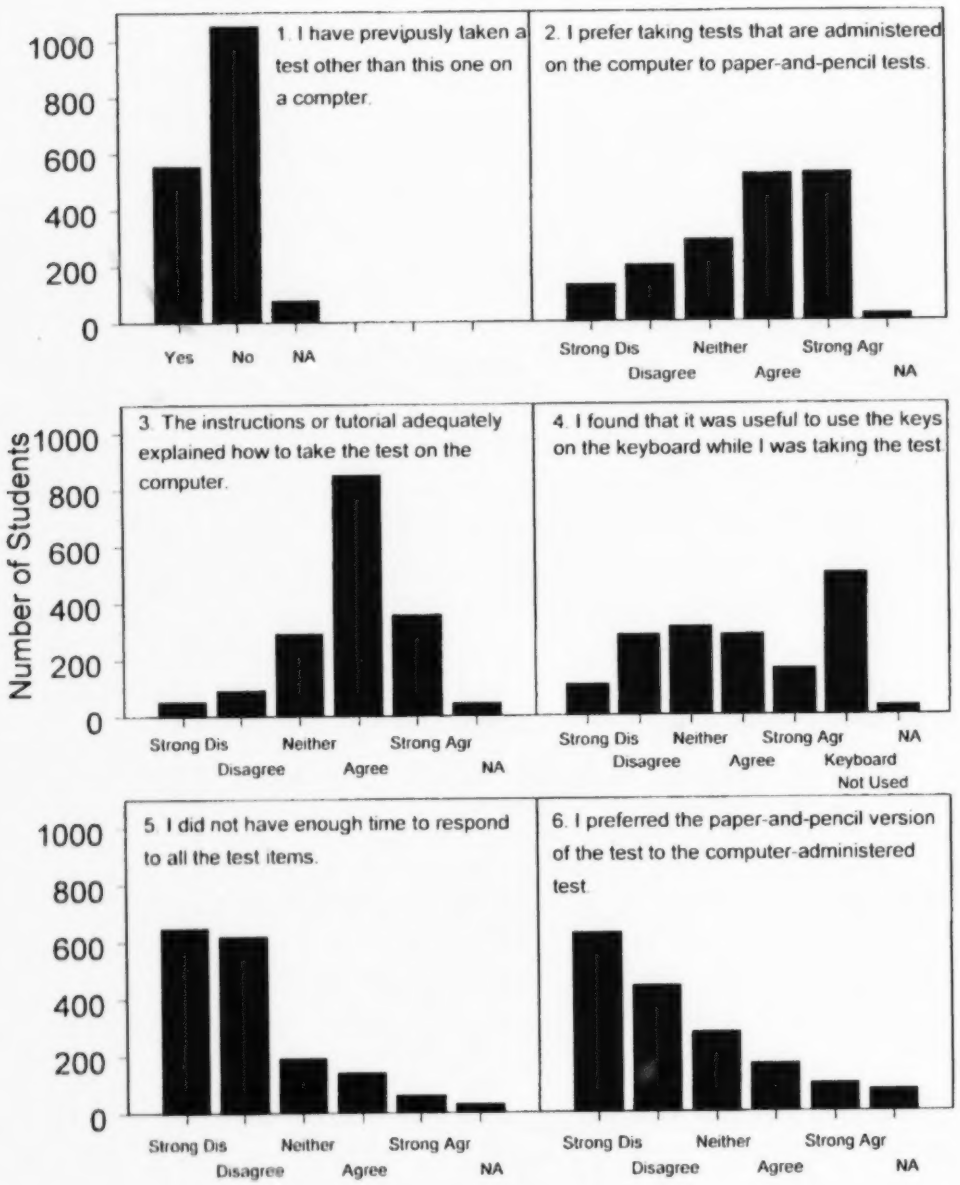

Figure 7.
Responses to survey questions for InView Level 3

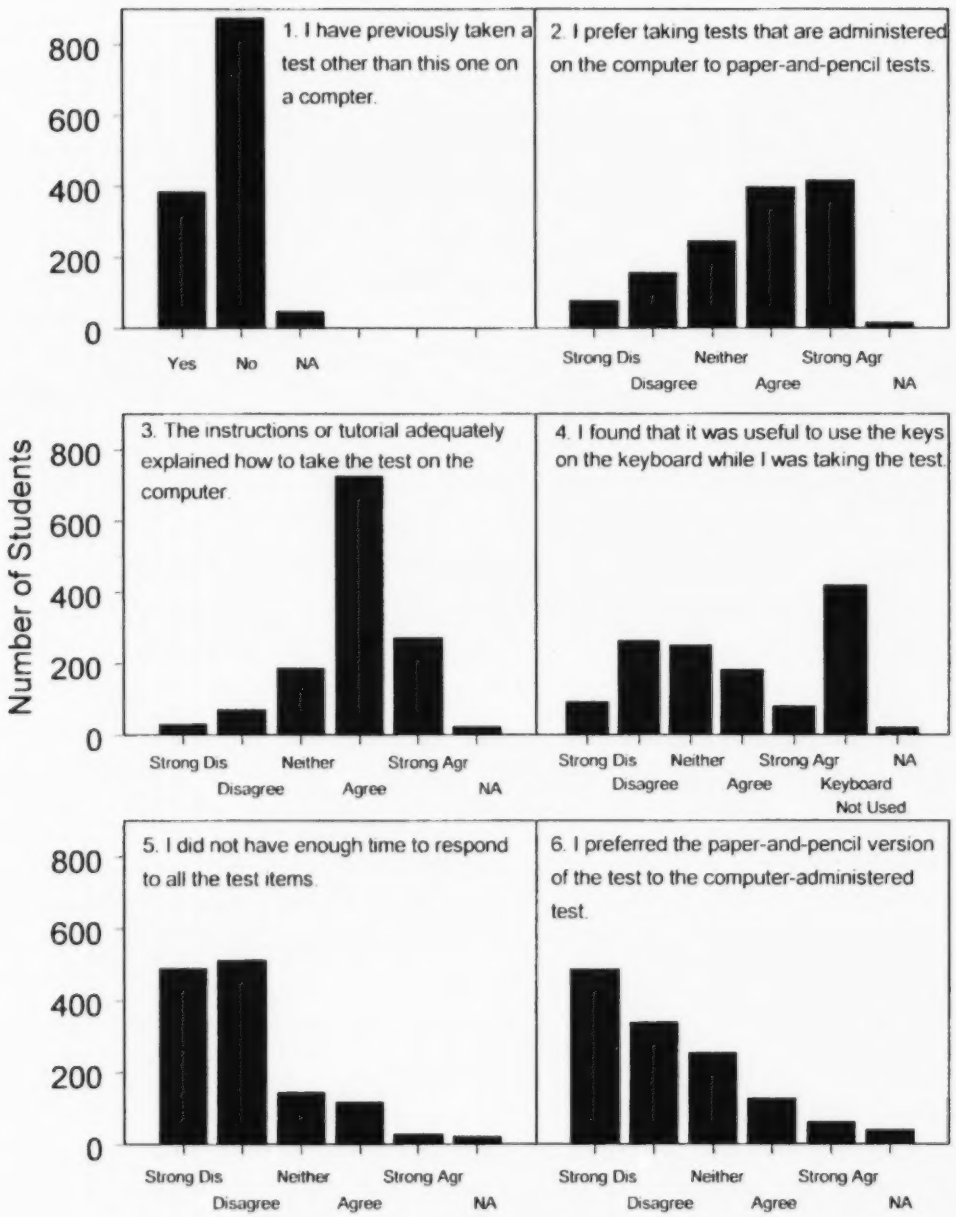
1. I have previously taken a test other than this one on a compter.
2. I prefer taking tests that are administered on the computer to paper-and-pencil tests.
3. The instructions or tutorial adequately explained how to take the test on the computer.
4. I found that it was useful to use the keys on the keyboard while I was taking the test.
5. I did not have enough time to respond to all the test items.
6. I preferred the paper-and-pencil version of the test to the computer-administered test.
Number of Students
0
200
400
600
800
Yes
No
NA
Strong Dis
Disagree
Neither
Agree
Strong Agr
NA
Keyboard Not Used

Figure 8.
Responses to survey questions for InView Level 4

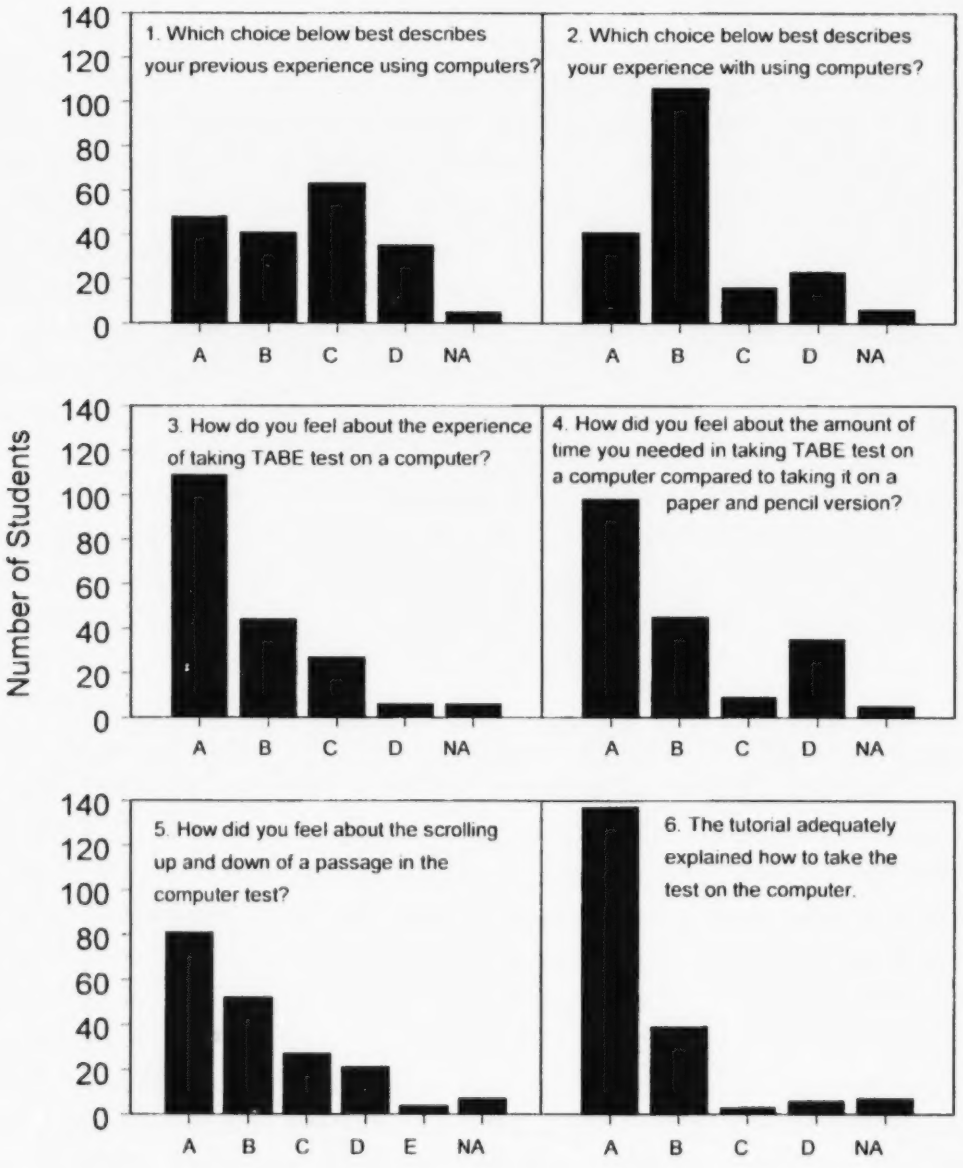
Number of Students
140
120
100
80
60
40
20
0
1. Which choice below best describes your previous experience using computers?
A B C D NA
2. Which choice below best describes your experience with using computers?
A B C D NA
3. How do you feel about the experience of taking TABE test on a computer?
A B C D NA
4. How did you feel about the amount of time you needed in taking TABE test on a computer compared to taking it on a paper and pencil version?
A B C D NA
5. How did you feel about the scrolling up and down of a passage in the computer test?
A B C D E NA
6. The tutorial adequately explained how to take the test on the computer.
A B C D NA

Figure 9.
Responses to survey questions for TABE Level M

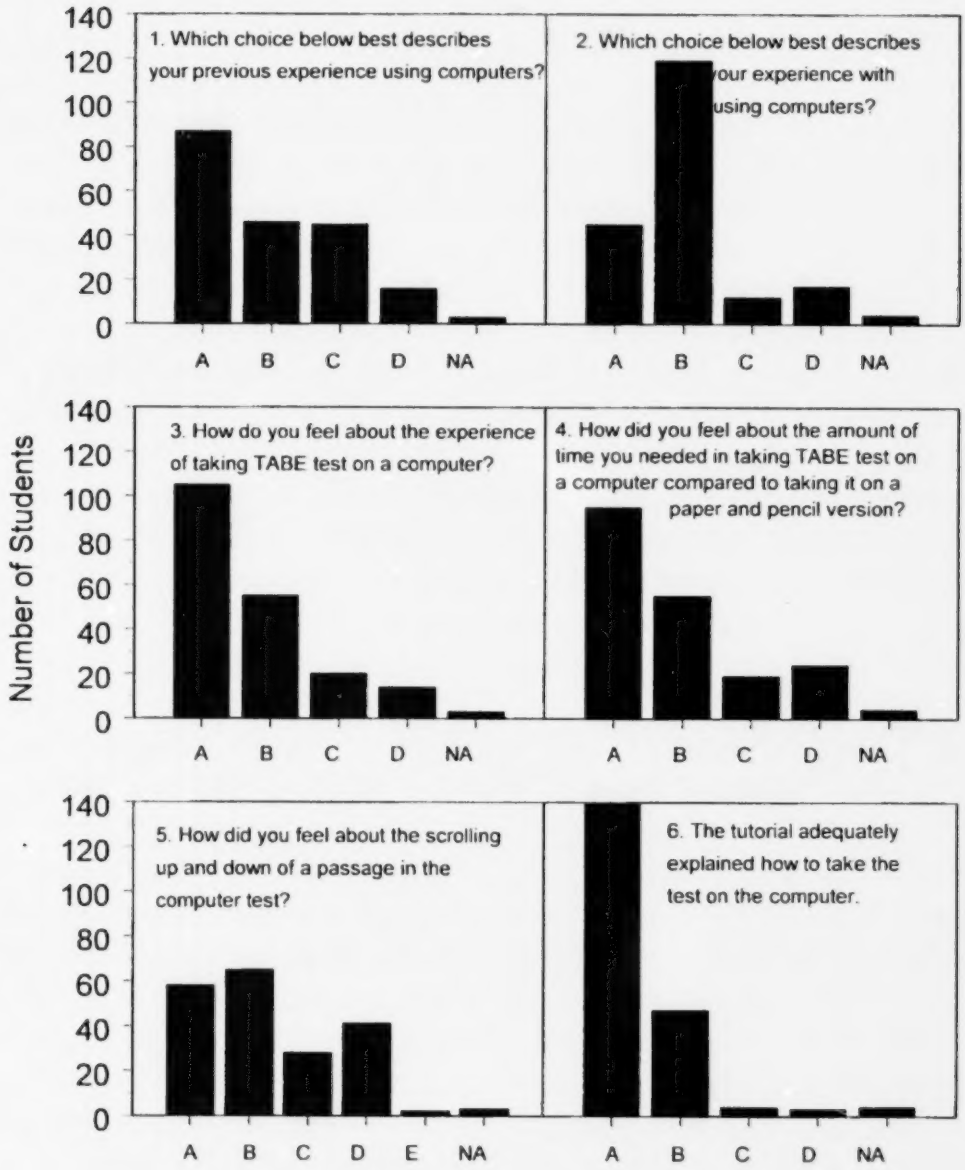
Number of Students
140
120
100
80
60
40
20
0
1. Which choice below best describes your previous experience using computers?
A B C D NA
2. Which choice below best describes
our experience with
using computers?
A B C D NA
3. How do you feel about the experience of taking TABE test on a computer?
A B C D NA
4. How did you feel about the amount of time you needed in taking TABE test on a computer compared to taking it on a paper and pencil version?
A B C D NA
5. How did you feel about the scrolling up and down of a passage in the computer test?
A B C D E NA
6. The tutorial adequately explained how to take the test on the computer.
A B C D NA

Figure 10.
Responses to survey questions for TABE Level D

Appendix A - Student Surveys

InView Computer Experience Survey

1) I have previously taken a test other that this one on a computer.
   1: Yes
   2: No

If you answered "NO" Go to Question #3.

2) I prefer taking tests that are administered on the computer to paper-and-pencil tests.
   1: Strongly Disagree
   2: Disagree
   3: Neither Agree nor Disagree
   4: Agree
   5: Strongly Agree

3) The instructions or tutorial adequately explained how to take the test on the computer.
   1: Strongly Disagree
   2: Disagree
   3: Neither Agree nor Disagree
   4: Agree
   5: Strongly Agree

4) I found that it was useful to use the keys on the keyboard while I was taking the test.
   1: Strongly Disagree
   2: Disagree
   3: Neither Agree nor Disagree
   4: Agree
   5: Strongly Agree
   6: Did not use the keyboard during the test

5) I did not have enough time to respond to all the test items.
   1: Strongly Disagree
   2: Disagree
   3: Neither Agree nor Disagree
   4: Agree
   5: Strongly Agree

6) I preferred the paper-and-pencil version of the test to the computer-administered test.
   1: Strongly Disagree
   2: Disagree
   3: Neither Agree nor Disagree
   4: Agree
   5: Strongly Agree

TABE On-line/PC Survey

This On-line/PC Survey addresses the following issues: accessibility/familiarity with computers, anxiety/comfort level with computers, speededness, scrolling effect, tutorial effectiveness.

1. Which choice below best describes your previous experience using computers?

   a. I use a computer often at home and at school/work.
   b. I use a computer often only at school or at work.
   c. I do not use computer very often.
   d. I rarely or never use computers.

2. Which choice below best describes your experience with using computers?

   a. I feel I do my best work on a computer.
   b. I am comfortable using a computer.
   c. Using computers does not appeal to me.
   d. I get anxious when I use a computer.

3. How do you feel about the experience of taking TABE test on a computer?

   a. It was easier than a paper and pencil test.
   b. It was about the same as taking a paper and pencil test.
   c. It was more difficult at first, but became easier as I got used to it.
   d. It was harder than taking a paper and pencil test.

4. How did you feel about the amount of time you needed in taking TABE test on a computer compared to taking it on a paper and pencil version?

   a. It took me less time to complete the test.
   b. It took me about the same amount of time to complete the test.
   c. It took me more time to complete the test.
   d. I wasn't really aware of how long the computer test was taking.

5. How did you feel about the scrolling up and down of a passage in the computer test?

   a. It was easier than reading it on paper and pencil test.
   b. It was about the same as reading it from a paper and pencil test.
   c. It was more difficult at first, but became easier as I got used to it.
   d. It was more difficult than reading it from a paper and pencil test.
   e. It did not occur on the test that I took.

6. The tutorial adequately explained how to take the test on the computer.
   a. Yes, very well.
   b. Yes, fairly well.

c. No, not very well.
d. No, not at all.

Appendix B

Derivations of Linn-Harnisch and the Standard Mean Difference Procedures

Linn-Harnisch. The Linn-Harnisch procedure uses the systematic differences between the obtained and expected frequencies derived via the three-parameter model . First, the sample is divided into ten equal score categories (deciles) based upon their location on the ability score (θ) scale for a given item. The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group. The proportion of people in decile g who are expected to answer item i correctly is

$$P_{ig} = \frac{1}{n_g} \sum_{j \in g} P_{ij}, \tag{1}$$

where n_g is the number of examinees in decile g. The proportion of people expected to answer item i correctly (over all deciles) for a group (e.g., students who use a scientific calculator) is:

$$P_{i\cdot} = \frac{\sum_{g=1}^{10} n_g P_{ig}}{\sum_{g=1}^{10} n_g} \tag{2}$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is the number of examinees in decile g who answered item i correctly divided by the number of students in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \in g} u_{ij}}{n_g}, \tag{3}$$

where u_{ij} is the dichotomous score for item i for examinee j.

The corresponding formula to compute the observed proportion, over all deciles, of students answering each item correctly in the group is given by:

$$O_{i\cdot} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}, \tag{4}$$

After the values are calculated for these variables, the difference between the observed proportion correct and expected proportion correct for a particular group can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item i in decile g is

$$D_{ig} = O_{ig} - P_{ig\cdot}, \tag{5}$$

and the overall group difference (D_{i}) between observed and expected proportion correct for item i in the complete group (over all deciles) is

$$D_{i\cdot} = O_{i\cdot} - P_{i\cdot}. \tag{6}$$

These indices are indicators of the degree to which members of a group perform better or worse than expected on each item, based on the parameters estimated from all groups. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference.

Items are flagged as demonstrating DIF for (+) or against (*) the specified subgroup according to the following rule: An item demonstrates DIF against a subgroup if the $D_{i.} \leq -0.10$ and $Z \leq 2.58$. DIF in favor of a subgroup is defined in the same way but with a positive difference.

Standardized Mean Difference. The Standardized Mean Difference (SMD) is an extension of the Mantel-Haenszel (MH) statistic used for calculating DIF where

$$SMD = \Sigma\, p_{Fk}\, m_{Fk} - \Sigma\, p_{Fk}\, m_{Rk}, \tag{7}$$

where

$$p_{Fk} = n_{F+k} / n_{F++} \tag{8}$$

is the proportion of focal group members who are at the k^{th} level of the matching variable,

$$m_{Fk} = (1/n_{F+k})(\Sigma\, y_i\, n_{Rik}) \tag{9}$$

is the mean item score for the focal group at the k^{th} level, and

$$m_{Rk} = (1/n_{R+k})(\Sigma\, y_i\, n_{Rik}) \tag{10}$$

is the analogous value for the reference group. A positive value for a SMD reflects DIF in favor of the focal group. Likewise, a negative SMD reflects DIF against the focal group.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035028

I. DOCUMENT IDENTIFICATION:

Title: A DIF Analysis of Item-Level Mode Effects for Computerized and Paper-and-Pencil Tests	
Author(s): Richard D. Schwarz, Changhua Rich, Tracy Podrabsky	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: S.	Printed Name/Position/Title: Changhua Rich, Ph.D. Research Scientist
Organization/Address: CTB/McGraw-Hill 20 Ryan Ranch Road Monterey, CA 93940	Telephone: 831 393 7417 FAX: 831 393 7016
	E-Mail Address: CRich@ctb.com Date: 6/2/2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)